DOCUMENT RESUME

ED 331 719 SE 052 017

AUTHOR Marshall, James E.
TITLE Construct Validity of Multiple-choice and Performance-based Assessments of Basic Science Process Skills: A Multitrait-Multimethod Analysis
PUB DATE 91
NOTE 14p.; Paper presented at the Annual Meeting of the National Assocation for Research in Science Teaching (Lake Geneva, WI, April 7-10, 1991).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Construct Validity; Educational Assessment; Formative Evaluation; Grade 7; Junior High Schools; Middle Schools; *Multiple Choice Tests; *Multitrait Multimethod Techniques; *Performance Tests; *Process Education; *Science Tests; Skill Development
IDENTIFIERS *Test of Basic Process Skills in Science

ABSTRACT

Science process skills are described as a set of broadly transferable abilities, appropriate to all of the science disciplines and reflective of the true behavior of scientists. While science process skills have gained wide acceptance as an integral part of the science curricula, the development of valid and reliable instruments to assess those skills has lagged behind. The purpose of this study was to gather evidence of the construct validity of the multiple-choice and performance-based versions of the Test of Basic Process Skills in Science (BAPS), the only research instrument designed to measure all of the most widely accepted basic science process skills for elementary and middle school students. A multitrait-multimethod construct validation technique was used to gather evidence of the convergent and discriminant validity of the BAPS tests. The BAPS multiple-choice test and the BAPS station test, a performance-based instrument, were used to measure the trait of interest. The Test of Logical Thinking (TOLT) and the Bending Rods (RODS) Piagetian manipulative task were used to measure the discriminant trait, science reasoning ability. The four instruments were administered to a sample of 151 seventh grade students from a west Florida school district. The results indicated strong support for the convergent and discriminant validity of the BAPS instrument. Considerable evidence of the construct validity of the BAPS tests can be inferred from this study. (Author/KR)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED331719

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

JAMES E. MARSHALL

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Construct Validity of Multiple-choice and Performance-based Assessments of Basic Science Process Skills: A Multitrait-Multimethod Analysis

By

James E. Marshall, Ph.D.
California State University, Fresno
School of Education
Fresno, CA 93740-0002

BEST COPY AVAILABLE

A paper presented at the annual meeting of the National Association for Research in Science Teaching, April 7-10, 1991, Lake Geneva, WI.

SE052017

Construct Validity of Multiple-choice and Performance-based Assessments of Basic Science Process Skills: A Multitrait-Multimethod Analysis

An Abstract

Science process skills are described as a set of broadly transferable abilities, appropriate to all of the science disciplines and reflective of the true behavior of scientists. While science process skills have gained wide acceptance as an integral part of science curricula over the past 25 years, the development of valid and reliable instruments to assess those skills has lagged behind. Presently, few research instruments exist which measure science process skills, particularly the basic science process skills. Of the existing tests, few are adequately validated. Therefore, the purpose of the study was to gather evidence of the construct validity of the multiple-choice and performance-based versions of the *Test of Basic Process Skills in Science* (BAPS), the only research instrument designed to measure all of the most widely accepted basic science process skills for elementary and middle school students.

A multitrait-multimethod construct validation technique was used to gather evidence of the convergent and discriminant validity of the BAPS tests. The technique requires the administration of at least two instruments, with different test formats, to measure a trait of interest, and at least two instruments, again with different formats, to measure a different or discriminant trait. The BAPS multiple-choice test and the BAPS station test, a performance-based instrument, were used to measure the trait of interest, basic science process skills. The *Test of Logical Thinking* (TOLT) and the *Bending Rods* (RODS) Piagetian manipulative task were used to measure the discriminant trait, science reasoning ability.

The four instruments were administered to a sample of 151 seventh grade students from a west Florida school district. The scores were then correlated and presented in a multitrait-multimethod matrix. Convergent validity was established by correlating the BAPS multiple-choice scores and the BAPS station

scores. The resulting correlation was .80 ($p < .001$), indicating strong evidence of convergent validity. Discriminant validity was established by correlating the BAPS tests scores with the reasoning ability scores. The resulting correlations were .28 between the BAPS multiple-choice and the TOLT and .31 between the BAPS multiple-choice and the RODS. Similarly low correlations were found between the BAPS station test and the TOLT (.36) and the RODS (.38) lending support for discriminant validity.

The results indicated strong support for the convergent and discriminant validity of the BAPS instruments. Considerable evidence of the construct validity of the BAPS tests can be inferred from this study. Implications of these results are discussed in regards to the current trend towards performance-based, authentic assessment in science.

BEST COPY AVAILABLE

Construct Validity of Multiple-choice and Performance-based Assessments of Basic Science Process Skills: A Multitrait-Multimethod Analysis

With the current emphasis in education on scientific literacy, improving ways to teach and measure science process skills has become a national priority. This study was designed to contribute to this improvement effort by gathering evidence of the construct validity of measures of basic science process skills. The evidence is useful to educators and researchers in understanding the construct of basic science process skills, and to states, districts, and classroom teachers in developing the science test instruments of the future.

Science process skills are described as a set of broadly transferable abilities, appropriate to all of the science disciplines and reflective of the true behavior of scientists. While science process skills have gained wide acceptance as an integral part of science curricula over the past 25 years, the development of valid and reliable instruments to assess those skills has lagged behind. A review of literature revealed a lack of instruments designed specifically to assess basic science process skills--the skills which provide the foundation for learning the more complex integrated process skills. In fact, only one instrument existed which purports to assess all of the basic science process skills--the *Test of Basic Process Skills in Science* (Padilla, Cronin, and Twiest, 1985). The purpose of the study was to gather evidence of the construct validity of the multiple-choice and performance-based versions of the *Test of Basic Process Skills in Science* (BAPS), and, in doing so, begin to develop an empirical base for further decision making in science assessment.

Methods

A multitrait-multimethod (MTMM) construct validation technique was used to gather evidence of the convergent and discriminant validity of the BAPS tests. Campbell and Fiske (1959) stated that the demonstration of construct validity requires both

convergent and discriminant validity; that is, multiple measures of the same construct should be substantially correlated with each other, but less correlated with measures of other constructs. The technique requires the administration of at least two instruments with different test formats, to measure a trait of interest, and at least two instruments, again with different formats, to measure a different or discriminant trait. Resulting scores are correlated and organized into a multitrait-multimethod matrix. Evidence of construct validity comes from an analysis of the degree to which measures of the same trait involving different methods correlate higher with each other than they do with measures of different traits involving different methods (Campbell and Fiske, 1959). Messick (1989) describes the MTMM matrix as a heuristic device, not an analytical procedure. But, it is a "tough (often humbling) heuristic device that forces the investigator to confront simultaneously both convergent and discriminant evidence, or the lack thereof."

Instrumentation

The BAPS multiple-choice test and the BAPS station test, a performance-based instrument, were used to measure the trait of interest, basic science process skills. Created for elementary and middle school students, the BAPS multiple-choice test emphasizes the basic process skills of observation, classification, communication, inference, prediction, and measurement. The authors used a four alternative, multiple-choice format with numerous pictures and diagrams. The test was rated content valid by a panel of experts. According to Padilla (1989), results of the initial validation were "confusing and disappointing." Subsequently, Twiest and Twiest (1989) revised the BAPS multiple-choice test and also developed a performance-based, station test to parallel the objectives of the multiple-choice instrument. It was this revised version of the BAPS multiple-choice test and the newly constructed BAPS station test (BAPSST) that were used in the present study.

The *Test of Logical Thinking* (TOLT) (Tobin and Capie, 1981) and the *Bending Rods* (RODS) (Inhelder and Piaget, 1958) Piagetian manipulative task were used to measure the discriminant trait,

science reasoning ability. Both the TOLT, a ten-item double multiple-choice test, and the RODS, a single performance task, have been used extensively in investigating science reasoning abilities and exhibit evidence of reliability and validity.

Sampling, Data Gathering and Analysis

The four instruments were administered to a sample of 151 seventh- grade students from a west Florida school district. Seventh-grade students were selected because they represent the age group in which the suggested age ranges of the four instruments overlap. The sample was stratified to insure that low (21.2%), average (45%), and high (23.8%) ability students and both male (51%) and female (49%) students were represented in the study.

The instruments were all administered with a one-week period. To insure that the order in which the tests were given did not bias the results, the admininstration was counterbalanced.

The instruments were handscored and the data computer analyzed. Because of the influence of skewness on correlations, frequency distributions of the scores were analyzed and appropriate transformations computed. The scores were then correlated and presented in a multitrait-multimethod matrix. Descriptive statistics and reliabilities were also computed and presented.

Results

Table 1 summarizes the descriptive statistics of the scores on the four instruments used in the study. The mean scores for the BAPS multiple-choice test and the BAPS station test were 28.03 and 22.75, respectively.

Insert Table 1 about here

Twiest and Twiest (1989) reported a similar difference between scores on multiple-choice instrument and the performance-based instrument. While the mean scores and ranges on the two tests of basic science process skills seem to indicate that the BAPS multiple-choice test is more difficult than the BAPS

station test, an exit interview of students revealed a conflicting perception. Prior to receiving their scores, 143 of the 151 students (94.7%) felt they had scored higher on the station test than on the multiple-choice test. In actuality, only six of the 151 students (4.0%) scored higher on the station test . The students almost unanamously reported that they would prefer to take a performance-based test rather than a multiple-choice test. Perception data was not collected following the reporting of scores to the students. More research in this area is suggested.

To determine the strength of the relationships among the scores on the four instruments, Pearson product-moment correlation coefficients were computed and arranged in a multitrait-multimethod matrix (Table 2). However, because the test scores were moderately skewed, square root transformations were computed prior to correlating the scores. All coefficients were positive and significantly different from zero, $p < .001$ (two-tailed test of significance).

Insert Table 2 about here

Reliability coefficients were placed along the leading diagonal as prescribed by the MTMM technique, and are shown in parentheses. Cronbach alpha coefficients of internal consistency were assessed for the BAPS, BAPSST, and the TOLT. The coefficients for the BAPS and the BAPSST were .84 and .85, respectively. The coefficient alpha for the TOLT was .90. Reliability for the RODS was assessed by correlating parallel forms of the instrument. The resulting coefficient was .92. While the MTMM technique is vulnerable to unreliability, these coefficients were sufficiently large and consistent enough to warrant their use in the MTMM analysis.

The classic Campbell and Fiske (1959) paper proposed a heuristic strategy for investigating the MTMM. This strategy involved simple inspection and intuitive comparison of the correlation indices but did not provide a formal assessment

framework. In order to stay close to the original philosophy of Campbell and Fiske, a general significance testing procedure was used to provide a formal framework for evaluating the indices. This procedure not only tested for the significance of correlations but also tested for significant differences between correlations.

Campbell and Fiske (1959) proposed criteria for analyzing a MTMM matrix. These criteria are: (1) convergent validities should be statistically significant and substantial; (2) convergent validities should be higher than the correlations between different traits assessed by different methods; and (3) convergent validities should be higher than the correlations between different traits assessed by the same methods. Criterion1 is related to convergent validity, while criteria 2 and 3 relate to discriminant validity. Therefore, the following analysis of the MTMM matrix is organized into two sections: convergent validity and discriminant validity.

Convergent validity

Convergent validity is indicated by significant and sufficiently large correlation between the two measures of the trait of interest (criterion 1). In this study, those two measures were the BAPS and the BAPSST. This correlation coefficient is found in the validity diagonal of the MTMM matrix (Table 2). This diagonal is represented on the matrix by the correlations in bold print. The convergent validity coefficient between the BAPS and the BAPSST was .80. This coefficient was statistically significant ($p < .001$) and was sufficiently large to warrant further examination. This result provides strong support for the convergent validity of the BAPS. The validity diagonal also contains the convergent validity coefficient for reasoning ability, the discriminant trait. That coefficient was .70 ($p < .001$).

Discriminant validity

Discriminant validity is addressed by Campbell and Fiske's criterion 2 and criterion 3. Criterion 2 states that convergent validity coefficients should be higher than the correlations between different traits assessed by different methods. These correlations in the MTMM matrix are enclosed by broken lines. The coefficient of .31 represents the correlation between the RODS test and the BAPS,

while the coefficient of .36 represents the correlation between the BAPSST and the TOLT. While both of these coefficients were statistically different from zero ($p < .001$), they only explained 9.6% and 12.9%, respectively, of the shared variance.

A *t* test was performed to assess the difference between these coefficients and the convergent validity coefficient for the BAPS. The test revealed a significant difference between the convergent validity coefficient of .80 and the values of .31 ($t(148) = 8.397$, $p < .05$) and .36 ($t(148) = 7.008$, $p < .05$). By fulfilling the requirements of criterion 2, these results lend support for the discriminant validity of the BAPS.

Criterion 3 states that convergent validity coefficients should be higher than the correlations between different traits assessed by the same methods. In the MTMM matrix, these correlations are enclosed in solid lines. The coefficient .28 represents the correlation between the BAPS and the TOLT, while the coefficient of .38 represents the correlation between the BAPSST and the RODS test. As with criterion 2, these coefficients were both statistically different from zero ($p < .001$), though only explaining 7.8% and 14.4%, respectively, of the shared variance.

Again, a *t* test was used to assess the difference between these coefficients and the convergent validity coefficient for the BAPS. The test revealed a statistically significant difference between the convergent validity coefficient of 0.80 and the values of .28 ($t(148) = 8.822$, $p < .05$) and .38 ($t(148) = 6.808$, $p < .05$). As with criterion 2, these results lend strong support for the discriminant validity of the BAPS.

Conclusions

The results indicated strong support for the convergent and discriminant validity of the BAPS instruments. Therefore, considerable evidence of the construct validity of the BAPS tests can be inferred from this study.

These findings have direct implications to the current trend toward performance-based assessment. On one hand, the results provide much needed empirical evidence of the potential for valid performance assessment in science education. However, on the

other hand, the relatively high correlation between the BAPS multiple-choice test and the BAPS performance-based test (r = .80) suggests that the more efficient and less costly multiple-choice instrument can assess many of the same processes as the performance instrument. What those processes are was not determined by this study. A follow-up factor analytic comparison is currently being conducted to further investigate the relationship between these two instruments and to better understand what it is that they are assessing.

This study in no way contradicts the logical contention that science assessment instruments should be aligned with the currently advocated instructional strategies. Acknowledging the potent ability of state-wide assessment programs to drive curriculum and instruction, promoting performance testing of science is essential. However, promoting performance testing over multiple-choice testing because of empirical differences is not supported by this study.

References

Campbell, D. and Fiske, D. (1959). Convergent and discriminant validity by multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

Inhelder, B. and Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* NY: Basic Books.

Messick, S. (1989). Validity. In, R. Linn, Ed. *Educational Measurement*, 3rd Ed. NY: MacMillan.

Padilla, M. (1989). *Assessment of higher order thinking in science.* A paper presented at the annual meeting of the National Association for Research in Science Teaching, San Francisco.

Padilla, M., Cronin, L., and Twiest, M. (1985). *The development and validation of a test of basic process skills.* A paper presented at the annual meeting of the National Association for Research in Science Teaching, French Lick, IN.

Tobin, K. and Capie, W. (1981). Development and validation of a group test of logical thinking. *Educational & Pyschological Measurement*, 41(2), 413-423.

Twiest, M. and Twiest, M. (1989). *Modification and validation of a test of basic process skills using different methods of administration.* A paper presented at the annual meeting of the National Association for Research in Science Teaching, San Francisco.

BEST COPY AVAILABLE

Table 1. Summary Table of Descriptive Statistics.

Instrument	Mean	Std Dev	Minimum	Maximum
BAPS Multiple-choice	28.03	4.56	11	35
BAPS Station Test	22.75	5.00	6	31
TOLT	1.83	2.01	0	8
RODS	0.61	0.99	0	3

Table 2. Multitrait-multimethod Matrix of Two Measures of Basic Process Skills (PS) and Two Measures of Reasoning Ability (RA)

		Paper and Pencil		Performance	
		BAPS (PS)	TOLT (RA)	BAPSST (PS)	RODS (RA)
Paper/pencil	BAPS	(.84)			
	TOLT	.28	(.90)		
Performance	BAPSST	**.80**	.36	(.85)	
	RODS	.31	**.70**	.38	(.92)

Note. The validity diagonal is formed by the values in bold print. The reliability diagonal is formed by the values in parentheses. The heterotrait-monomethod values are enclosed by solid lines, while the heterotrait-heteromethod values are enclosed by broken lines. For the purposes of the MTMM technique, the trait factors (basic science process skills and reasoning ability) are positioned within the method factors (paper and pencil and performance).
Note. All coefficients are significant at the 0.001 level.
n = 151